Proc. of a Satellite Workshop of LREC 2008 on Corpora for Research on Emotion and Affect
Marrakesh, 2008

Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus

A. Batliner, S. Steidl, E. Nöth

Chair of Pattern Recognition, Friedrich-Alexander-University Erlangen-Nuremberg (FAU), Germany
email: {batliner,steidl,noeth}@informatik.uni-erlangen.de

Abstract

We report on a thoroughly processed and annotated German emotional speech database (children interacting with Sony's Aibo robot): 51 children, some 48 k words, 9.2 hours of speech, 5 labellers, word-based annotation of emotional user states. Several additional annotations as well as a mapping onto higher units of different granularity have been carried out. The database will eventually be made available for scientific use; in the licensing agreement, we plan to include mandatory benchmark constellations in order to make a comparison across sites possible.

1. Introduction[1]

Even if the terminology has not been standardised yet – there is no agreement as for the *exact* meaning of 'naturalistic', 'realistic', 'spontaneous', etc. – it is generally agreed upon that non-acted emotional databases should be aimed at. This might not be mandatory for generic, basic research but holds especially if we think of any application that eventually, outside of the laboratory, has to deal with non-acted data – simply because classifiers have to be trained with data that are as close as possible to the 'real' data. Obviously, the effort needed for designing, recording, and annotating spontaneous emotional databases is way higher than the one needed for acted data; thus, some acted databases are (freely) available such as the Berlin Database of Emotional Speech 'Emo-DB' (Burkhardt et al., 2005) or the Danish Emotional Speech Database 'DES' (Engberg et al., 1997) but, at least to our knowledge, no spontaneous one, at least not on similar conditions. Moreover, privacy reasons often prevent such data to be released to third parties. Thus, access to spontaneous data is the most severe bottleneck for a 'realistic' processing and emotion classification.

In the years 2002-2004, we have collected and processed a spontaneous emotional German database at FAU Erlangen within the EU-project PF-STAR. In the years 2005-2007, this database has been further annotated, and processed outside and within the so-called CEICES initiative (Batliner et al., 2006) within the NoE HUMAINE. There exist several publications on experiments using this database; however, its description has always been rather short, concentrating on those aspects that we focused on in the respective papers; this was simply due to the usual space restriction. As we eventually decided to release the database for scientific use, in this paper we want to give a condensed overview of aspects, annotations, and conditions of use. After a general description of the design and the recordings, we will give an account of the annotations which will be made available. In the end, we decided to call the database the 'FAU Aibo Emotion Corpus' (abbr.: FAU Aibo) because there exist other 'Aibo' corpora with emotional speech, cf. (Tato et al., 2002; Küstner et al., 2004).

Of course, we do not conceive the strategies chosen and presented in this paper as the only and best ones but as reasonable choices. At the end of most (sub-)sections, we will motivate our choices and partly discuss them against the backdrop of possible alternative solutions. These comments will be given in italics.

2. Material

The general frame for FAU Aibo is human–robot communication, children's speech, and the elicitation and subsequent recognition of emotional user states. The robot is Sony's (dog-like) robot AIBO. The basic idea is to combine a rather so far neglected type of corpus (children's speech) with 'natural' emotional speech within a Wizard-of-Oz task. The children were not told to use specific instructions but to talk to the Aibo like they would talk to a friend. They were led to believe that the Aibo is responding to their commands, but the robot is actually being controlled by a human operator, using the 'Aibo Navigator' software over a wireless LAN (the existing Aibo speech recognition module is not used). The wizard causes the Aibo to perform a fixed, pre-determined sequence of actions, which takes no account of what the child says. For the sequence of Aibo's actions, we tried to find a good compromise between obedient and disobedient behaviour: we wanted to provoke the children in order to elicit emotional behaviour but of course we did not want to run the risk that they break off the experiment. The children believed that the Aibo was reacting to their orders - albeit often not immediately. In fact, it was the other way round: the Aibo always strictly followed the same screen-plot, and the children had to align their orders to it's actions.

The data was collected from 51 children (age 10 - 13, 21 male, 30 female). The children were from two different schools, Mont and Ohm; the recordings took place in the resp. class-rooms. Speech was transmitted with a wireless head set (UT 14/20 TP SHURE UHF-series with microphone WH20TQG) and recorded with a DAT-recorder. The sampling rate of the signals is 48 kHz, quantisation is 16 bit. The data is downsampled to 16 kHz. Each record-

[1] This work was funded by the EU in the projects PF-STAR under grant IST-2001-37599 and HUMAINE under grant IST-2002-50742. The database has been further processed in the initiative CEICES within HUMAINE (**C**ombining **E**fforts for **I**mproving automatic **C**lassification of **E**motional user **S**tates). The responsibility for the contents of this study lies with the authors.

ing session took some 30 minutes. The speech data were segmented automatically into speech files ('turns'), triggering a turn boundary at pauses $\geq$ 1.5 seconds. Note that here, the term 'turn' does not imply any linguistic meaning; however, it turned out that only in very few cases, this criterion wrongly decided in favour of a turn boundary instead of (implicitly) modelling a hesitation pause. Because of the experimental setup, these recordings contain a huge amount of silence (reaction time of the Aibo), which caused a noticeable reduction of recorded speech after raw segmentation; finally we obtained about 9.2 hours of speech.

Children as early adapters within an edu-/entertainment scenario should be plausible addressees for automatic emotion modelling. Sometimes it has been doubted that they are 'representative' because they might behave unlike adults. Of course, speech recognition and feature extraction have to be adapted slightly – the same way as procedures designed only for male speakers have to be adapted for female speakers. Of course, the children can display group-specific tendencies but there is no indication that they behaved differently from adults in a fundamental way.

3. Further Processing

3.1. Transliteration and word lexica

The orthographic transliteration – in pared down VERBMOBIL notation – was done by advanced students and cross-checked by the supervisor. The phonetic word lexicon is in SAMPA notation. In addition, we established a syntactic-semantic word lexicon, with coarse part-of-speech (POS) labels per word (six classes), and with coarse higher semantic labels per word (six classes modelling valence and some other word types such as vocative).

While modelling linguistic information, normally words are not used as such but processed somehow – at least they are stemmatised for, e.g., bag-of-word modelling. In our experience, such a very coarse mapping onto six POS or higher semantic classes only still yields a pretty good classification performance on the spoken word chain. Of course, there are many other mappings which can be conceived. However, our 'simple' classes can be used as benchmark for other approaches, cf. below.

3.2. Word-based emotion annotation

In other studies, the unit of analysis is normally given trivially – a read sentence, a dialogue move, etc. – or defined intuitively. We conceive the word as the smallest possible emotional unit; even if we cannot exclude the possibility of changing emotions within the same word, this will definitely be a rather exotic exception. By annotating word-based, we are later on free to map words onto longer units. Our strategy thus allows to find 'optimal' units of analysis on an empirical basis.

Five labellers (advanced students of linguistics, 4 females, 1 male) listened to the speech files in sequential order and annotated independently from each other each word as neutral (default) or as belonging to one of ten other classes which were obtained by inspection of the data; we do not claim that they represent children's emotions in general, only that they are adequate for the modelling of the behaviour of these children in this specific scenario. We resort to majority voting (henceforth MV): if three or more labellers agree, the label is attributed to the word; if four or five labellers agree, we assume some sort of prototypes. The following raw labels were used; in parentheses, the number of cases with MV is given: *joyful* (101), *surprised* (0), *emphatic* (2528), *helpless* (3), *touchy*, i. e., irritated (225), *angry* (84), *motherese* (1260), *bored* (11), *reprimanding* (310), *rest*, i. e. non-neutral, but not belonging to the other categories (3), *neutral* (39169); 4707 words had no MV; all in all, there were 48401 words. *joyful* and *angry* belong to the 'big' emotions, the other ones rather to 'emotion-related/emotion-prone' user states. The state *emphatic* has been introduced because it can be seen as a possible indication of some (starting) trouble in communication and by that, as a sort of 'pre-emotional' state (Batliner et al., 2005; Batliner et al., 2008).

A single database is no omnibus in the sense that choosing a specific scenario for the recordings pre-defines the range of classes one can observe; what cannot be observed cannot be modelled. However, we claim that our data are fairly representative for realistic data: only a few of the 'classic', big n emotions, and a very skewed distribution. Instead, one 'emotion-related' state comes on the scene, i. e. motherese. Emphatic is, in fact, just a possible pre-stage of emotion. However, in some form it will often be observed in such realistic settings; thus it makes sense to model it.

3.3. Partitioning into sub-samples

Some of the labels are very sparse; if we only take labels with more than 50 MVs, this 7-class problem is most interesting from a methodological point of view, cf. the new dimensional representation of these seven category labels in (Batliner et al., 2008). However, the distribution of classes is very unbalanced. Therefore, we downsampled *neutral* and *emphatic* and mapped *touchy* and *reprimanding*, together with *angry*, onto **A***ngry*[2] as representing different but closely related kinds of negative attitude. For this more balanced 4-class problem AMEN, 1557 words for **A***ngry* (**A**), 1224 words for **M***otherese* (**M**), and 1645 words each for **E***mphatic* (**E**),and for **N***eutral* (**N**), are used, cf. (Steidl et al., 2005). Cases where less than three labellers agreed were omitted as well as those cases where other than these four main classes were labelled. For this AMEN subset, weighted kappa is 0.59.

This sub-sample has been used in several experiments and will be defined as (the basis of) the main 'canonical' samples to be processed, cf. below.

3.4. Chunking into and mapping of labels onto syntactically and emotionally meaningful units

Now we were facing the task of mapping word-based labels onto higher units, first onto turns: a simple 50% threshold – for instance, if an **A** turn has 10 words , then 5 or more

[2] If we refer to the resulting 4-class problem, the initial letter is given boldfaced and recte. Note that now, **A***ngry* can consist, for instance, of two *touchy* and one *reprimanding* label; thus the number of **A***ngry* cases is far higher than the sum of *touchy*, *reprimanding*, and *angry* MV cases.

words have to be labelled as **A** – would be suboptimal because some words, esp. function words, are likely not to be produced in an emotional manner; moreover, a longer turn can consist of one neutral clause, and one emotional clause - then chances are that the whole turn will wrongly be mapped onto neutral.

For the mapping onto turn-based labels, we employed the following strategy: as stop words, fragments and auxiliaries were used; for the turns containing our 6070 AMEN words, this means 17618 words, 3996 turns; stopwords are: 596 fragments, 196 auxiliaries (some words both), i. e. 16856 words remaining.[3] For each turn, we add together the labels given by our 5 labellers (for *n* words, 5 x *n* labels). If the turn is mapped onto neutral, 70% of the labels have to be neutral. (*joyful* and the other spurious labels are not taken into account for this computing.) If 30% or more are non-neutral, then the turn is **A**, **M**, or **E**. If at least 50% of the non-neutral labels are **M**, the turn is mapped onto **M**. If **A** and **E** are equally distributed, the turn is mapped onto **A**. If the turn is neither **A** nor **M**, it is **E**. This simply means that we employ a sort of 'markedness' condition: **M** is more marked than **A**, and **A** is more marked than **E**, and all are more marked than **N**. This yields the following turn-based labels: 868 **A** (21.7 %), 1347 **E** (33.7 %), 495 **M** (12.4 %), and 1280 **N** (32.0 %), summing up to 3990 (100 %) labels = turns.

For the mapping onto 'chunk-based' labels in between word level and turn level, we first annotated the whole database with a coarse syntactic boundary system (main/subordinate clauses, free phrases, dislocations, and vocatives as label especially tuned for these data); we then used similar mapping rules as for the turns. The rules are given explicitly in a structogram in a forthcoming paper.

Our 'turns' are similar to the units used in other studies. As they can consist of up to 53 words, they are not really optimal – we claim that our chunks are. By using different thresholds etc., the chunk size can be adapted to specific needs; the same way, different chunk sizes can be established for finding out how classifiers behave if faced with shorter or longer units. A pivotal characteristic of this solution is that our chunks are syntactically – and by that, semantically – well defined. This is a necessary prerequisite for higher linguistic (deep or shallow) processing in any end-to-end automatic dialogue system.

3.5. Automatic forced alignment per word plus manual correction of this automatic segmentation

We did an automatic forced alignment using the transliteration (i. e., the spoken word chain). Such an alignment is nowadays rather good but of course sometimes erroneous. We therefore decided to have the automatic word segmentation corrected manually for the whole database; the segmentation of the 3990 AMEN turns was cross-checked by the first author.

A corrected reference segmentation allows to exclude wrong segmentation as a source of misclassification, and makes comparisons across approaches more reliable because at least this factor can be kept constant.

3.6. Automatic pitch extraction plus manual correction of this extraction

Historically, pitch has had a prominent position w. r. t. all feature types because of the preponderance of intonation models in the last decades. Even if this might not be mirrored in empirical results, it is of course an important parameter which is, however, notoriously known as impossible to be extracted fully reliably. Databases with corrected pitch values are rare, and we do not know of any other emotion database with such corrected values. We therefore decided to use, in addition to our own pitch detection algorithm (PDA), a well-known frame-based PDA as baseline and correct these values manually. This was done for the 3990 AMEN turns by the first author. More details and differences in classification performance can be found in (Batliner et al., 2007b; Batliner et al., 2007a).

Such manually corrected pitch values do not constitute a ground truth; they are of course biased towards the automatic PDA used. A pitch-synchronous correction was not possible, due to time constraints. Note that even 'objective' measures such as laryngographic recordings are no ground truth: they are close to the signal but not close to perception! However, the corrected values can be used for computing F0 features and be compared to such features based on – sometimes erroneous – automatic PDAs. Again, this helps in keeping constant at least one factor, namely the raw pitch values, and makes comparisons across different approaches of computing pitch features more reliable.

3.7. Further annotations, software, and types of data

In dialogues, the dialogue partner's reaction can be valuable information that can be coded and used in classification. In our scenario, the Aibo does not speak and has no facial gestures. However, we can model its behaviour - whether it is co-operative or not. It seems to be plausible that a non-co-operative behaviour triggers negative reactions to a larger extent than co-operative behaviour. Therefore, we annotated the Aibo's [± co-operative] actions, although, because of the effort needed, only for roughly half of the data. This annotation will not be part of the default distribution but can be made available on a bilateral basis.

In connection with FAU Aibo, two software programs were made available to the community: EDE - Evaluating Decoders using Entropy, and eLabel - Labelling of Emotions.[4] Within the CEICES initiative, a feature encoding scheme has been developed aiming at a full coverage of possible acoustic and linguistic features (low level descriptors and functionals). It is ASCII-based but could easily be converted into some other (e.g., XML) representation. This encoding scheme will be made available on demand.

Apart from the close-talk microphone recordings, there are two more types of recordings: one with the microphone of the video-camera used for protocolling the sessions containing noise and reverberation, and a second one which is

[3] Note that of course, we could find some more stop words, but this would be rather data-driven and not generic so we refrained from that.

[4] http://www5.informatik.uni-erlangen.de/en/our-team/steidl-stefan/free-software-in-humaine/

artificially reverberated. These are not part of the distribution but can be made available on a bilateral basis. (Note that for privacy reasons, the video recordings will not be available.) More details are given in (Schuller et al., 2007).

4. Mandatory benchmarking

Apart from the lack of (freely) available spontaneous, realistic databases in the field of emotion classification – and because of this lack – comparisons of performance across studies, let alone strict evaluations using the very same data such as the ones conducted within the NIST initiative, are practically impossible. Note that even in the case of rather well-defined acted databases such as Emo-DB and DES (well-defined because the classes are given trivially via speaker instructions), researchers often do not select exactly the same cases per class. This makes a comparison of performance across studies impossible. However, in our experience, a convenient selection of sub-samples out from a whole corpus often contributes more to classification performance than the choice of the one or the other feature selection procedure or classifier. Even if this selection is well documented (which is not always the case), it is not clear how much it contributes, if there is no baseline setting. Thus for our database, we want to define and make available in the distribution an 'evaluation setting': a simple two-fold cross-validation which can be computed with rather low effort. As benchmark, we will provide classification results for this constellation which is defined extensionally and thus unambiguous; its processing – in addition to any other processing – will be mandatory.

Of course, the data can and should be exploited in many different ways. Experience tells us, however, that often, data are chosen in a way that results in highest possible recognition rates – and this is what readers remember. By defining obligatory constellations which are, on the same time, simple enough and do not need a high processing effort, we want to establish a sort of benchmark which has to be used by licencees in studies on these data.

5. Concluding Remarks

Using our annotations, the impact of alternative approaches can be pursued (different size of units of analysis, automatically vs. manually extracted values, etc.); seen from a performance point of view, the difference might not be marked; however, adding several small differences might yield larger ones.

The field of automatic classification of emotional user states is still in its infancy, compared to other fields such as automatic speech recognition. This is foremost due to its topic: whereas a word is a word is a word, even in noisy condition, it is neither clear what an emotion is, nor where we can find which emotion in which unit of analysis. All this has to be annotated somehow. This makes it expensive to create databases[5], and databases greater by some order of magnitude would be one prerequisite for standardization in this field. What we hopefully can aim at in the next time to come is thus not a strict evaluations but something like 'islands of standardization': studies dealing, for example, with FAU Aibo can be compared w.r.t. the benchmark constellation.

6. References

A. Batliner, S. Steidl, C. Hacker, E. Nöth, and H. Niemann. 2005. Tales of Tuning – Prototyping for Automatic Classification of Emotional User States. In *Proc. 9th Eurospeech - Interspeech 2005*, pages 489–492, Lisbon.

A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson. 2006. Combining Efforts for Improving Automatic Classification of Emotional User States. In *Proceedings of IS-LTC 2006*, pages 240–245, Ljubliana.

A. Batliner, S. Steidl, and E. Nöth. 2007a. Laryngealizations and Emotions: How Many Babushkas? In *Proceedings of the International Workshop on Paralinguistic Speech — between Models and Data (ParaLing'07)*, pages 17–22, Saarbrücken.

A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson. 2007b. The Impact of F0 Extraction Errors on the Classification of Prominence and Emotion. In *Proceedings of ICPhS 2007*, pages 2201–2204, Saarbrücken.

A. Batliner, S. Steidl, C. Hacker, and E. Nöth. 2008. Private emotions vs. social interaction — a data-driven approach towards analysing emotions in speech. *User Modeling and User-Adapted Interaction*, 18:175–206.

F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. 2005. A database of german emotional speech. In *Proc. 9th Eurospeech - Interspeech 2005*, pages 1517–1520, Lisbon.

Inger S. Engberg, Anya Varnich Hansen, Ove Andersen, and Paul Dalsgaard. 1997. Design, recording and verification of a Danish emotional speech database. In *Proc. Eurospeech*, pages 1695–1698, Rhodes.

D. Küstner, R. Tato, T. Kemp, and B. Meffert. 2004. Towards Real Life Applications in Emotion Recognition. In E. André, L. Dybkiaer, W. Minker, and P. Heisterkamp, editors, *Affective Dialogue Systems*, pages 25–35, Berlin, Springer.

B. Schuller, D. Seppi, A. Batliner, A. Meier, and S. Steidl. 2007. Towards more Reality in the Recognition of Emotional Speech. In *Proc. of ICASSP 2007*, pages 941–944, Honolulu.

S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann. 2005. "Of All Things the Measure is Man": Automatic Classification of Emotions and Inter-Labeler Consistency. In *Proc. of ICASSP 2005*, pages 317–320, Philadelphia.

R. Tato, R. Santos, R. Kompe, and J.M. Pardo. 2002. Emotional space Improves Emotion Recognition. In *Proc. ICSLP 2002*, pages 2029–2032.

[5] A (very!) coarse estimate of our expenses for designing, recording, and processing manually FAU Aibo amounts to $>$ 80 K Euros for researchers and students.